DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model

Shohei Shimizu*, Takanori Inazumi†, Yasuhiro Sogawa‡,
Aapo Hyvärinen§, Yoshinobu Kawahara¶, Takashi Washio‖
Patrik O. Hoyer**, Kenneth Bollen††

Abstract

Structural equation models and Bayesian networks have been widely used to analyze causal relations between continuous variables. In such frameworks, linear acyclic models are typically used to model the data-generating process of variables. Recently, it was shown that use of non-Gaussianity identifies the full structure of a linear acyclic model, *i.e.*, a causal ordering of variables and their connection strengths, without using any prior knowledge on the network structure, which is not the case with conventional methods. However, existing estimation methods are based on iterative search algorithms and may not converge to a correct solution in a finite number of steps. In this paper, we propose a new direct method to estimate a causal ordering and connection strengths based on non-Gaussianity. In contrast to the previous methods, our algorithm requires no algorithmic parameters and is guaranteed to converge to the right solution within a small fixed number of steps if the data strictly follows the model.

*The Institute of Scientific and Industrial Research (ISIR), Osaka University, Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan. Email: sshimizu@ar.sanken.osaka-u.ac.jp

†The Institute of Scientific and Industrial Research (ISIR), Osaka University, Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan. Email: inazumi@ar.sanken.osaka-u.ac.jp

‡The Institute of Scientific and Industrial Research (ISIR), Osaka University, Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan. Email: sogawa@ar.sanken.osaka-u.ac.jp

§Department of Computer Science, Department of Mathematics and Statistics and Helsinki Institute for Information Technology, University of Helsinki, FIN-00014, Finland. Email: aapo.hyvarinen@helsinki.fi

¶The Institute of Scientific and Industrial Research (ISIR), Osaka University, Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan. Email: kawahara@ar.sanken.osaka-u.ac.jp

‖The Institute of Scientific and Industrial Research (ISIR), Osaka University, Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan. Email: washio@ar.sanken.osaka-u.ac.jp

**Department of Computer Science, University of Helsinki, FIN-00014, Finland. Email: patrik.hoyer@helsinki.fi

††Department of Sociology, CB 3210 Hamilton Hall, University of North Carolina, Chapel Hill, NC 27599-3210, U.S.A. Email: bollen@unc.edu

1 Introduction

Many empirical sciences aim to discover and understand causal mechanisms underlying various natural phenomena and human social behavior. An effective way to study causal relationships is to conduct a controlled experiment. However, performing controlled experiments is often ethically impossible or too expensive in many fields including social sciences [1], bioinformatics [2] and neuroinformatics [3]. Thus, it is necessary and important to develop methods for causal inference based on the data that do not come from such controlled experiments.

Structural equation models (SEM) [1] and Bayesian networks (BN) [4,5] are widely applied to analyze causal relationships in many empirical studies. A *linear* acyclic model that is a special case of SEM and BN is typically used to analyze causal effects between continuous variables. Estimation of the model commonly uses only the covariance structure of the data and in most cases cannot identify the full structure, *i.e.*, a causal ordering and connection strengths, of the model with no prior knowledge on the structure [4, 5].

In [6], a non-Gaussian variant of SEM and BN called a linear non-Gaussian acyclic model (LiNGAM) was proposed, and its full structure was shown to be identifiable without pre-specifying a causal order of the variables. This feature is a significant advantage over the conventional methods [4, 5]. A non-Gaussian method to estimate the new model was also developed in [6] and is closely related to independent component analysis (ICA) [7]. In the subsequent studies, the non-Gaussian framework has been extended in various directions for learning a wider variety of SEM and BN [8–10]. In what follows, we refer to the non-Gaussian model as LiNGAM and the estimation method as ICA-LiNGAM algorithm.

Most of major ICA algorithms including [11,12] are iterative search methods [7]. Therefore, the ICA-LiNGAM algorithms based on the ICA algorithms need some additional information including initial guess and convergence criteria. Gradient-based methods [11] further need step sizes. However, such algorithmic parameters are hard to optimize in a systematic way. Thus, the ICA-based algorithms may get stuck in local optima and may not converge to a reasonable solution if the initial guess is badly chosen [13].

In this paper, we propose a new direct method to estimate a causal ordering of variables in the LiNGAM without prior knowledge on the structure. The new method estimates a causal order of variables by successively reducing each independent component from given data in the model, and this process is completed in steps equal to the number of the variables in the model. It is not based on iterative search in the *parameter space* and needs no initial guess or similar algorithmic parameters. It is *guaranteed* to converge to the right solution within a small fixed number of steps if the data *strictly* follows the model, *i.e.*, if all the model assumptions are met and the sample size is infinite. These features of the new method enable more accurate estimation of a causal order of the variables in a disambiguated and direct procedure. Once the causal orders of variables is identified, the connection strengths between the variables are easily estimated

using some conventional covariance-based methods such as least squares and maximum likelihood approaches [1]. We also show how prior knowledge on the structure can be incorporated in the new method.

The paper is structured as follows. First, in Section 2, we briefly review LiNGAM and the ICA-based LiNGAM algorithm. We then in Section 3 introduce a new direct method. The performance of the new method is examined by experiments on artificial data in Section 4, and experiments on real-world data in Section 5. Conclusions are given in Section 6. Preliminary results were presented in [14–16].

2 Background

2.1 A linear non-Gaussian acyclic model: LiNGAM

In [6], a non-Gaussian variant of SEM and BN, which is called LiNGAM, was proposed. Assume that observed data are generated from a process represented graphically by a directed acyclic graph, *i.e.*, DAG. Let us represent this DAG by a $m \times m$ adjacency matrix $\mathbf{B}=\{b_{ij}\}$ where every b_{ij} represents the connection strength from a variable x_j to another x_i in the DAG. Moreover, let us denote by $k(i)$ a causal order of variables x_i in the DAG so that no later variable determines or has a directed path on any earlier variable. (A directed path from x_i to x_j is a sequence of directed edges such that x_j is reachable from x_i.) We further assume that the relations between variables are linear. Without loss of generality, each observed variable x_i is assumed to have zero mean. Then we have

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i, \tag{1}$$

where e_i is an external influence. All external influences e_i are continuous random variables having *non-Gaussian* distributions with zero means and non-zero variances, and e_i are independent of each other so that there is no latent confounding variables [5].

We rewrite the model (1) in a matrix form as follows:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{2}$$

where $\mathbf{x}$ is a p-dimensional random vector, and $\mathbf{B}$ could be permuted by simultaneous equal row and column permutations to be *strictly* lower triangular due to the acyclicity assumption [1]. Strict lower triangularity is here defined as a lower triangular structure with all zeros on the diagonal. Our goal is to estimate the adjacency matrix $\mathbf{B}$ by observing data $\mathbf{x}$ only. Note that we do not assume that the distribution of $\mathbf{x}$ is faithful [5] to the generating graph.

We note that each b_{ij} represents the direct causal effect of x_j on x_i and each a_{ij}, the (i, j)-the element of the matrix $\mathbf{A}=(\mathbf{I}-\mathbf{B})^{-1}$, the total causal effect of x_j on x_i [17].

We emphasize that x_i is equal to e_i if no other observed variable x_j $(j \neq i)$ inside the model has a directed edge to x_i, *i.e.*, all the b_{ij} $(j \neq i)$ are zeros. In

such a case, an external influence e_i is *observed* as x_i. Such an x_i is called an *exogenous observed* variable. Otherwise, e_i is called an *error*. For example, consider the model defined by

$$\begin{aligned} x_2 &= e_2 \\ x_1 &= 1.5x_2 + e_1 \\ x_3 &= 0.8x_1 - 1.5x_2 + e_3, \end{aligned}$$

where x_2 is equal to e_2 since it is not determined by either x_1 or x_3. Thus, x_2 is an exogenous observed variable, and e_1 and e_3 are errors. Note that there *exists at least one exogenous observed variable* $x_i(=e_i)$ due to the acyclicity and the assumption of no latent confounders.

An exogenous observed variable is usually defined as an observed variable that is determined outside of the model [1]. In other words, an exogenous observed variable is a variable that any other observed variable inside the model does not have a directed edge to. The definition does not require that it is equal to an independent external influence, and the external influences of exogenous observed variables may be dependent. However, in the LiNGAM (2), an exogenous observed variable is always equal to an independent external influence due to the assumption of no latent confounders.

2.2 Identifiability of the model

We next explain how the connection strengths of the LiNGAM (2) can be identified as shown in [6]. Let us first solve Eq. (2) for $\mathbf{x}$. Then we obtain

$$\mathbf{x} = \mathbf{A}\mathbf{e}, \tag{3}$$

where $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$ is a mixing matrix whose elements are called mixing coefficients and can be permuted to be lower triangular as well due to the aforementioned feature of $\mathbf{B}$ and the nature of matrix inversion. Since the components of $\mathbf{e}$ are independent and non-Gaussian, Eq. (3) defines the independent component analysis (ICA) model [7], which is known to be identifiable [18, 19].

ICA essentially can estimate $\mathbf{A}$ (and $\mathbf{W} = \mathbf{A}^{-1} = \mathbf{I} - \mathbf{B}$), but has permutation, scaling and sign indeterminacies. ICA actually gives $\mathbf{W}_{ICA}$=$\mathbf{PDW}$, where $\mathbf{P}$ is an unknown permutation matrix, and $\mathbf{D}$ is an unknown diagonal matrix. But in LiNGAM, the correct permutation matrix $\mathbf{P}$ can be found [6]: the correct $\mathbf{P}$ is the only one that gives no zeros in the diagonal of $\mathbf{DW}$ since $\mathbf{B}$ should be a matrix that can be permuted to be strictly lower triangular and $\mathbf{W} = \mathbf{I} - \mathbf{B}$. Further, one can find the correct scaling and signs of the independent components by using the unity on the diagonal of $\mathbf{W}$=$\mathbf{I}$−$\mathbf{B}$. One only has to divide the rows of $\mathbf{DW}$ by its corresponding diagonal elements to obtain $\mathbf{W}$. Finally, one can compute the connection strength matrix $\mathbf{B} = \mathbf{I} - \mathbf{W}$.

2.3 ICA-LiNGAM algorithm

The ICA-LiNGAM algorithm presented in [6] is described as follows:

ICA-LiNGAM algorithm

1. Given a p-dimensional random vector $\mathbf{x}$ and its $p \times n$ observed data matrix $\mathbf{X}$, apply an ICA algorithm (FastICA using hyperbolic tangent function [12]) to obtain an estimate of $\mathbf{A}$.
2. Find the unique permutation of rows of $\mathbf{W}=\mathbf{A}^{-1}$ which yields a matrix $\widetilde{\mathbf{W}}$ without any zeros on the main diagonal. The permutation is sought by minimizing $\sum_i 1/|\widetilde{\mathbf{W}}_{ii}|$.
3. Divide each row of $\widetilde{\mathbf{W}}$ by its corresponding diagonal element, to yield a new matrix $\widetilde{\mathbf{W}}'$ with all ones on the diagonal.
4. Compute an estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$ using $\widehat{\mathbf{B}} = \mathbf{I} - \widetilde{\mathbf{W}}'$.
5. Finally, to estimate a causal order $k(i)$, find the permutation matrix $\widetilde{\mathbf{P}}$ of $\widehat{\mathbf{B}}$ yielding a matrix $\widetilde{\mathbf{B}} = \widetilde{\mathbf{P}}\widehat{\mathbf{B}}\widetilde{\mathbf{P}}^T$ which is as close as possible to a strictly lower triangular structure. The lower-triangularity of $\widetilde{\mathbf{B}}$ can be measured using the sum of squared b_{ij} in its upper triangular part $\sum_{i \leq j} \widetilde{b}_{ij}^2$ for small number of variables, say less than 8. For higher-dimensional data, the following approximate algorithm is used, which sets small absolute valued elements in $\widetilde{\mathbf{B}}$ to zero and tests if the resulting matrix is possible to be permuted to be strictly lower triangular:
   (a) Set the $p(p+1)/2$ smallest (in absolute value) elements of $\widehat{\mathbf{B}}$ to zero.
   (b) Repeat
      i. Test if $\widehat{\mathbf{B}}$ can be permuted to be strictly lower triangular. If the answer is yes, stop and return the permuted $\widehat{\mathbf{B}}$, that is, $\widetilde{\mathbf{B}}$.
      ii. Additionally set the next smallest (in absolute value) element of $\widehat{\mathbf{B}}$ to zero.

2.4 Potential problems of ICA-LiNGAM

The original ICA-LiNGAM algorithm has several potential problems: i) Most ICA algorithms including FastICA [12] and gradient-based algorithms [11] may not converge to a correct solution in a finite number of steps if the initially guessed state is badly chosen [13] or if the step size is not suitably selected for those gradient-based methods. The appropriate selection of such algorithmic parameters is not easy. In contrast, our algorithm proposed in the next section is guaranteed to converge to the right solution in a fixed number of steps equal to the number of variables if the data *strictly* follows the model. ii) The permutation algorithms in Steps 2 and 5 are not scale-invariant. Hence they could give a different or *even wrong* ordering of variables depending on scales or standard deviations of variables especially when they have a wide range of scales. However, scales are essentially not relevant to the ordering of variables.

Though such bias would vanish for large enough sample sizes, for practical sample sizes, an estimated ordering could be affected when variables are normalized to make unit variance for example, and hence the estimation of a causal ordering becomes quite difficult.

3 A direct method: DirectLiNGAM

3.1 Identification of an exogenous variable based on non-Gaussianity and independence

In this subsection, we present two lemmas and a corollary[1] that ensure the validity of our algorithm proposed in the next subsection 3.2. The basic idea of our method is as follows. We first find an exogenous variable based on its independence of the residuals of a number of pairwise regressions (Lemma 1). Next, we remove the effect of the exogenous variable from the other variables using least squares regression. Then, we show that a LiNGAM also holds for the residuals (Lemma 2) and that the same ordering of the residuals is a causal ordering for the original observed variables as well (Corollary 1). Therefore, we can find the second variable in the causal ordering of the original observed variables by analyzing the residuals and their LiNGAM, *i.e.*, by applying Lemma 1 to the residuals and finding an "exogenous" residual. The iteration of these effect removal and causal ordering estimates the causal order of the original variables.

We first quote Darmois-Skitovitch theorem [20, 21] since it is used to prove Lemma 1:

Theorem 1 (Darmois-Skitovitch theorem) *Define two random variables y_1 and y_2 as linear combinations of independent random variables s_i(i=1, $\cdots$, q):*

$$y_1 = \sum_{i=1}^{q} \alpha_i s_i, \quad y_2 = \sum_{i=1}^{q} \beta_i s_i. \tag{4}$$

Then, if y_1 and y_2 are independent, all variables s_j for which $\alpha_j \beta_j \neq 0$ are Gaussian. □

In other words, this theorem means that if there exists a non-Gaussian s_j for which $\alpha_j \beta_j \neq 0$, y_1 and y_2 are dependent.

Lemma 1 *Assume that the input data* $\mathbf{x}$ *strictly follows the LiNGAM (2). Denote by $r_i^{(j)}$ the residuals when x_i are regressed on x_j: $r_i^{(j)} = x_i - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)} x_j$ $(i \neq j)$. Then a variable x_j is exogenous if and only if x_j is independent of its residuals $r_i^{(j)}$ for all $i \neq j$.* □

[1] We prove the lemmas and corollary without assuming the faithfulness [5] unlike our previous work [14].

Proof (i) Assume that x_j is exogenous, *i.e.*, $x_j{=}e_j$. Due to the model assumption and Eq. (3), one can write $x_i{=}a_{ij}x_j{+}\bar{e}_i^{(j)}$ $(i{\neq}j)$, where $\bar{e}_i^{(j)}{=}\sum_{h\neq j} a_{ih}e_h$ and x_j are independent, and a_{ij} is a mixing coefficient from x_j to x_i in Eq. (3). The mixing coefficient a_{ij} is equal to the regression coefficient when x_i is regressed on x_j since $\mathrm{cov}(x_i, x_j){=}a_{ij}\mathrm{var}(x_j)$. Thus, the residual $r_i^{(j)}$ is equal to the corresponding error term, *i.e.*, $r_i^{(j)}{=}\bar{e}_i^{(j)}$. This implies that x_j and $r_i^{(j)}(=\bar{e}_i^{(j)})$ are independent.

(ii) Assume that x_j is not exogenous, *i.e.*, x_j has at least one parent. Let P_j denote the (non-empty) set of the variable subscripts of parent variables of x_j. Then one can write $x_j = \sum_{h\in P_j} b_{jh}x_h + e_j$, where x_h and e_j are independent and each b_{jh} is non-zero. Let a vector $\mathbf{x}_{P_j}$ and a column vector $\mathbf{b}_{P_j}$ collect all the variables in P_j and the corresponding connection strengths, respectively. Then, the covariances between $\mathbf{x}_{P_j}$ and x_j are

$$\begin{aligned} E(\mathbf{x}_{P_j}x_j) &= E\{\mathbf{x}_{P_j}(\mathbf{b}_{P_j}^T\mathbf{x}_{P_j} + e_j)\} \\ &= E(\mathbf{x}_{P_j}\mathbf{b}_{P_j}^T\mathbf{x}_{P_j}) + E(\mathbf{x}_{P_j}e_j) \\ &= E(\mathbf{x}_{P_j}\mathbf{x}_{P_j}^T)\mathbf{b}_{P_j}. \end{aligned} \tag{5}$$

The covariance matrix $E(\mathbf{x}_{P_j}\mathbf{x}_{P_j}^T)$ is positive definite since the external influences e_h that correspond to those parent variables x_h in P_j are mutually independent and have positive variances. Thus, the covariance vector $E(\mathbf{x}_{P_j}x_j) = E(\mathbf{x}_{P_j}\mathbf{x}_{P_j}^T)\mathbf{b}_{P_j}$ in Eq. (5) cannot equal the zero vector, and there must be at least one variable x_i $(i \in P_j)$ with which x_j covaries, *i.e.*, $\mathrm{cov}(x_i, x_j){\neq}0$. Then, for such a variable x_i $(i \in P_j)$ that $\mathrm{cov}(x_i, x_j){\neq}0$, we have

$$\begin{aligned} r_i^{(j)} &= x_i - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}x_j \\ &= x_i - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}\left(\sum_{h\in P_j} b_{jh}x_h + e_j\right) \\ &= \left\{1 - \frac{b_{ji}\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}\right\}x_i - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}\sum_{h\in P_j, h\neq i} b_{jh}x_h \\ &\quad - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}e_j. \end{aligned} \tag{6}$$

Each of those parent variables x_h (including x_i) in P_j is a linear combination of external influences *other than* e_j due to the relation of x_h to e_j that $x_j = \sum_{h\in P_j} b_{jh}x_h + e_j = \sum_{h\in P_j} b_{jh}\left(\sum_{k(t)\leq k(h)} a_{ht}e_t\right) + e_j$, where e_t and e_j are independent. Thus, the $r_i^{(j)}$ and x_j can be rewritten as linear combinations of

independent external influences as follows:

$$r_i^{(j)} = \left\{1 - \frac{b_{ji}\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}\right\}\left(\sum_{l\neq j} a_{il}e_l\right) - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}\sum_{h\in P_j, h\neq i} b_{jh}\left(\sum_{t\neq j} a_{ht}e_t\right) - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)}e_j. \quad (7)$$

$$x_j = \sum_{h\in P_j} b_{jh}\left(\sum_{t\neq j} a_{ht}e_t\right) + e_j. \quad (8)$$

The first two terms of Eq. (7) and the first term of Eq. (8) are linear combinations of external influences other than e_j, and the third term of Eq. (7) and the second term of Eq. (8) depend only on e_j and do not depend on the other external influences. Further, all the external influences including e_j are mutually independent, and the coefficient of non-Gaussian e_j on $r_i^{(j)}$ and that on x_j are non-zero. These imply that $r_i^{(j)}$ and x_j are dependent since $r_i^{(j)}$, x_j and e_j correspond to y_1, y_2, s_j in Darmois-Skitovitch theorem, respectively.

From (i) and (ii), the lemma is proven. ■

Lemma 2 *Assume that the input data* $\mathbf{x}$ *strictly follows the LiNGAM (2). Further, assume that a variable* x_j *is exogenous. Denote by* $\mathbf{r}^{(j)}$ *a (p-1)-dimensional vector that collects the residuals* $r_i^{(j)}$ *when all* x_i *of* $\mathbf{x}$ *are regressed on* x_j *($i\neq j$). Then a LiNGAM holds for the residual vector* $\mathbf{r}^{(j)}$*:* $\mathbf{r}^{(j)} = \mathbf{B}^{(j)}\mathbf{r}^{(j)} + \mathbf{e}^{(j)}$*, where* $\mathbf{B}^{(j)}$ *is a matrix that can be permuted to be strictly lower-triangular by a simultaneous row and column permutation, and elements of* $\mathbf{e}^{(j)}$ *are non-Gaussian and mutually independent.* □

Proof Without loss of generality, assume that $\mathbf{B}$ in the LiNGAM (2) is already permuted to be strictly lower triangular and that $x_j{=}x_1$. Note that $\mathbf{A}$ in Eq. (3) is also lower triangular (although its diagonal elements are all ones). Since x_1 is exogenous, a_{i1} are equal to the regression coefficients when x_i are regressed on x_1 ($i \neq 1$). Therefore, after removing the effects of x_1 from x_i by least squares estimation, one gets the first column of $\mathbf{A}$ to be a zero vector, and x_1 does not affect the residuals $r_i^{(1)}$. Thus, we again obtain a lower triangular mixing matrix $\mathbf{A}^{(1)}$ with all ones in the diagonal for the residual vector $\mathbf{r}^{(1)}$ and hence have a LiNGAM for the vector $\mathbf{r}^{(1)}$. ■

Corollary 1 *Assume that the input data* $\mathbf{x}$ *strictly follows the LiNGAM (2). Further, assume that a variable* x_j *is exogenous. Denote by* $k_{r^{(j)}}(i)$ *a causal order of* $r_i^{(j)}$*. Recall that* $k(i)$ *denotes a causal order of* x_i*. Then, the same ordering of the residuals is a causal ordering for the original observed variables as well:* $k_{r^{(j)}}(l){<}k_{r^{(j)}}(m) \Leftrightarrow k(l){<}k(m)$. □

Proof As shown in the proof of Lemma 2, when the effect of an exogenous variable x_1 is removed from the other observed variables, the second to p-th columns of $\mathbf{A}$ remain the same, and the submatrix of $\mathbf{A}$ formed by deleting the first row and the first column is still lower triangular. This shows that the ordering of the other variables is not changed and proves the corollary. ■

Lemma 2 indicates that the LiNGAM for the $(p-1)$-dimensional residual vector $\mathbf{r}^{(j)}$ can be handled as a new input model, and Lemma 1 can be further applied to the model to estimate the next exogenous variable (the next exogenous residual in fact). This process can be repeated until all variables are ordered, and the resulting order of the variable subscripts shows the causal order of the original observed variables according to Corollary 1.

To apply Lemma 1 in practice, we need to use a measure of independence which is not restricted to uncorrelatedness since least squares regression gives residuals always uncorrelated with but not necessarily independent of explanatory variables. A common independence measure between two variables y_1 and y_2 is their mutual information $MI(y_1, y_2)$ [7]. In [22], a nonparametric estimator of mutual information was developed using kernel methods.[2] Let K_1 and K_2 represent the Gram matrices whose elements are Gaussian kernel values of the sets of n observations of y_1 and y_2, respectively. The Gaussian kernel values $K_1(y_1^{(i)}, y_1^{(j)})$ and $K_2(y_2^{(i)}, y_2^{(j)})$ $(i, j = 1, \cdots, n)$ are computed by

$$K_1(y_1^{(i)}, y_1^{(j)}) = \exp\left(-\frac{1}{2\sigma^2}\|y_1^{(i)} - y_1^{(j)}\|^2\right) \tag{9}$$

$$K_2(y_2^{(i)}, y_2^{(j)}) = \exp\left(-\frac{1}{2\sigma^2}\|y_2^{(i)} - y_2^{(j)}\|^2\right), \tag{10}$$

where $\sigma>0$ is the bandwidth of Gaussian kernel. Further let κ denote a small positive constant. Then, in [22], the kernel-based estimator of mutual information is defined as:

$$\widehat{MI}_{kernel}(y_1, y_2) = -\frac{1}{2}\log\frac{\det \mathcal{K}_\kappa}{\det \mathcal{D}_\kappa}, \tag{11}$$

where

$$\mathcal{K}_\kappa = \begin{bmatrix} \left(K_1 + \frac{n\kappa}{2}I\right)^2 & K_1K_2 \\ K_2K_1 & \left(K_2 + \frac{n\kappa}{2}I\right)^2 \end{bmatrix} \tag{12}$$

$$\mathcal{D}_\kappa = \begin{bmatrix} \left(K_1 + \frac{n\kappa}{2}I\right)^2 & 0 \\ 0 & \left(K_2 + \frac{n\kappa}{2}I\right)^2 \end{bmatrix}. \tag{13}$$

As the bandwidth σ of Gaussian kernel tends to zero, the population counterpart of the estimator converges to the mutual information up to second order when it is expanded around distributions with two variables y_1 and y_2 being independent [22]. The determinants of the Gram matrices K_1 and K_2 can be efficiently

[2] Matlab codes can be downloaded at `http://www.di.ens.fr/~fbach/kernel-ica/index.htm`

computed by using the incomplete Cholesky decomposition to find their low-rank approximations of rank M ($\ll n$). In [22], it was suggested that the positive constant κ and the width of the Gaussian kernel σ are set to $\kappa = 2 \times 10^{-3}$, $\sigma = 1/2$ for $n > 1000$ and $\kappa = 2 \times 10^{-2}$, $\sigma = 1$ for $n \leq 1000$ due to some theoretical and computational considerations.

In this paper, we use the kernel-based independence measure. We first evaluate pairwise independence between a variable and each of the residuals and next take the sum of the pairwise measures over the residuals. Let us denote by U the set of the subscripts of variables x_i, *i.e.*, U=$\{1, \cdots, p\}$. We use the following statistic to evaluate independence between a variable x_j and its residuals $r_i^{(j)} = x_i - \frac{\mathrm{cov}(x_i,x_j)}{\mathrm{var}(x_j)} x_j$ when x_i is regressed on x_j:

$$T_{kernel}(x_j; U) \quad = \quad \sum_{i \in U, i \neq j} \widehat{MI}_{kernel}(x_j, r_i^{(j)}). \tag{14}$$

Many other nonparametric independence measures [23, 24] and more computationally simple measures that use a single nonlinear correlation [25] have also been proposed. Any such proposed method of independence could potentially be used instead of the kernel-based measure in Eq. (14).

3.2 DirectLiNGAM algorithm

We now propose a new direct algorithm called DirectLiNGAM to estimate a causal ordering and the connection strengths in the LiNGAM (2):

DirectLiNGAM algorithm

1. Given a p-dimensional random vector $\mathbf{x}$, a set of its variable subscripts U and a $p \times n$ data matrix of the random vector as $\mathbf{X}$, initialize an ordered list of variables $K := \emptyset$ and $m := 1$.

2. Repeat until $p-1$ subscripts are appended to K:

   (a) Perform least squares regressions of x_i on x_j for all $i \in U \backslash K$ $(i \neq j)$ and compute the residual vectors $\mathbf{r}^{(j)}$ and the residual data matrix $\mathbf{R}^{(j)}$ from the data matrix $\mathbf{X}$ for all $j \in U \backslash K$. Find a variable x_m that is most independent of its residuals:

   $$x_m = \arg \min_{j \in U \backslash K} T_{kernel}(x_j; U \backslash K), \tag{15}$$

   where T_{kernel} is the independence measure defined in Eq. (14).

   (b) Append m to the end of K.

   (c) Let $\mathbf{x} := \mathbf{r}^{(m)}$, $\mathbf{X} := \mathbf{R}^{(m)}$.

3. Append the remaining variable to the end of K.

4. Construct a strictly lower triangular matrix $\mathbf{B}$ by following the order in K, and estimate the connection strengths b_{ij} by using some conventional covariance-based regression such as least squares and maximum likelihood approaches on the original random vector $\mathbf{x}$ and the original data matrix $\mathbf{X}$. We use least squares regression in this paper.

3.3 Computational complexity

Here, we consider the computational complexity of DirectLiNGAM compared with the ICA-LiNGAM with respect to sample size n and number of variables p. A dominant part of DirectLiNGAM is to compute Eq. (14) for each x_j in Step 2(a). Since it requires $O(np^2M^2+p^3M^3)$ operations [22] in $p-1$ iterations, complexity of the step is $O(np^3M^2+p^4M^3)$, where M ($\ll n$) is the maximal rank found by the low-rank decomposition used in the kernel-based independence measure. Another dominant part is the regression to estimate the matrix $\mathbf{B}$ in Step 4. The complexity of many representative regressions including the least square algorithm is $O(np^3)$. Hence, we have a total budget of $O(np^3M^2+p^4M^3)$. Meanwhile, the ICA-LiNGAM requires $O(p^4)$ time to find a causal order in Step 5. Complexity of an iteration in FastICA procedure at Step 1 is known to be $O(np^2)$. Assuming a constant number C of the iterations in FastICA steps, the complexity of the ICA-LiNGAM is considered to be $O(Cnp^2+p^4)$. Though general evaluation of the required iteration number C is difficult, it can be conjectured to grow linearly with regards to p. Hence the complexity of the ICA-LiNGAM is presumed to be $O(np^3+p^4)$.

Thus, the computational cost of DirectLiNGAM would be larger than that of ICA-LiNGAM especially when the low-rank approximation of the Gram matrices is not so efficient, *i.e.*, M is large. However, we note the fact that DirectLiNGAM has guaranteed convergence in a fixed number of steps and is of known complexity, whereas for typical ICA algorithms including FastICA, the run-time complexity and the very convergence are not guaranteed.

3.4 Use of prior knowledge

Although DirectLiNGAM requires no prior knowledge on the structure, more efficient learning can be achieved if some prior knowledge on a part of the structure is available because then the number of causal orders and connection strengths to be estimated gets smaller.

We present three lemmas to utilize prior knowledge in DirectLiNGAM. Let us first define a matrix $\mathbf{A}^{\mathrm{knw}}=[a_{ji}^{\mathrm{knw}}]$ that collects prior knowledge under the LiNGAM (2) as follows:

$$a_{ji}^{\mathrm{knw}} := \begin{cases} 0 & \text{if } x_i \text{ does } \textit{not} \text{ have a directed path to } x_j \\ 1 & \text{if } x_i \text{ has a directed path to } x_j \\ -1 & \text{if no prior knowledge is available to know if either} \\ & \text{of the two cases above (0 or 1) is true.} \end{cases} \tag{16}$$

Due to the definition of exogenous variables and that of prior knowledge matrix $\mathbf{A}^{\mathrm{knw}}$, we readily obtain the following three lemmas.

Lemma 3 *Assume that the input data* $\mathbf{x}$ *strictly follows the LiNGAM (2). An observed variable* x_j *is exogenous if* a_{ji}^{knw} *is zero for all* $i \neq j$.

Lemma 4 *Assume that the input data* $\mathbf{x}$ *strictly follows the LiNGAM (2). An observed variable* x_j *is endogenous, i.e., not exogenous, if there exist such* $i \neq j$ *that* a_{ji}^{knw} *is unity.*

Lemma 5 *Assume that the input data* $\mathbf{x}$ *strictly follows the LiNGAM (2). An observed variable* x_j *does not receive the effect of* x_i *if* a_{ji}^{knw} *is zero.*

The principle of making DirectLiNGAM algorithm more accurate and faster based on prior knowledge is as follows. We first find an exogenous variable by applying Lemma 3 instead of Lemma 1 if an exogenous variable is identified based on prior knowledge. Then we do not have to evaluate independence between any observed variable and its residuals. If no exogenous variable is identified based on prior knowledge, we next find endogenous (non-exogenous) variables by applying Lemma 4. Since endogenous variables are never exogenous we can narrow down the search space to find an exogenous variable based on Lemma 1. We can further skip to compute the residual of an observed variable and take the variable itself as the residual if its regressor does not receive the effect of the variable due to Lemma 5. Thus, we can decrease the number of causal orders and connection strengths to be estimated, and it improves the accuracy and computational time. The principle can also be used to further analyze the residuals and find the next exogenous residual because of Corollary 1. To implement these ideas, we only have to replace Step 2a in DirectLiNGAM algorithm by the following steps:

2a-1 Find such a variable(s) x_j ($j \in U \backslash K$) that the j-th row of $\mathbf{A}^{\mathrm{knw}}$ has zero in the i-th column for all $i \in U \backslash K$ ($i \neq j$) and denote the set of such variables by U_{exo}. If U_{exo} is not empty, set $U_c := U_{exo}$. If U_{exo} is empty, find such a variable(s) x_j ($j \in U \backslash K$) that the j-th row of $\mathbf{A}^{\mathrm{knw}}$ has unity in the i-th column for at least one of $i \in U \backslash K$ ($i \neq j$), denote the set of such variables by U_{end} and set $U_c := U \backslash K \backslash U_{end}$.

2a-2 Denote by $V^{(j)}$ a set of such a variable subscript $i \in U \backslash K$ ($i \neq j$) that $a_{ij}^{\mathrm{knw}} = 0$ for all $j \in U_c$. First set $\mathbf{r}_i^{(j)} := \mathbf{x}_i$ for all $i \in V^{(j)}$, next perform least squares regressions of x_i on x_j for all $i \in U \backslash K \backslash V^{(j)}$ ($i \neq j$) and estimate the residual vectors $\mathbf{r}^{(j)}$ and the residual data matrix $\mathbf{R}^{(j)}$ from the data matrix $\mathbf{X}$ for all $j \in U_c$. If U_c has a single variable, set the variable to be x_m. Otherwise, find a variable x_m in U_c that is most independent of the residuals:

$$x_m = \arg\min_{j \in U_c} T_{kernel}(x_j; U \backslash K), \tag{17}$$

where T_{kernel} is the independence measure defined in Eq. (14).

4 Simulations

We first randomly generated 5 datasets based on sparse networks under each combination of number of variables p and sample size n (p=10, 20, 50, 100; n=500, 1000, 2000):

1. We constructed the $p \times p$ adjacency matrix with all zeros and replaced every element in the lower-triangular part by independent realizations of Bernoulli random variables with success probability s similarly to [26]. The probability s determines the sparseness of the model. The expected number of adjacent variables of each variable is given by $s(p-1)$. We randomly set the sparseness s so that the number of adjacent variables was 2 or 5 [26].

2. We replaced each non-zero (unity) entry in the adjacency matrix by a value randomly chosen from the interval $[-1.5, -0.5] \cup [0.5, 1.5]$ and selected variances of the external influences e_i from the interval $[1, 3]$ as in [27]. We used the resulting matrix as the data-generating adjacency matrix $\mathbf{B}$.

3. We generated data with sample size n by independently drawing the external influence variables e_i from various 18 non-Gaussian distributions used in [22] including (a) Student with 3 degrees of freedom; (b) double exponential; (c) uniform; (d) Student with 5 degrees of freedom; (e) exponential; (f) mixture of two double exponentials; (g)-(h)-(i) symmetric mixtures of two Gaussians: multimodal, transitional and unimodal; (j)-(k)-(l) nonsymmetric mixtures of two Gaussians, multimodal, transitional and unimodal; (m)-(n)-(o) symmetric mixtures of four Gaussians: multimodal, transitional and unimodal; (p)-(q)-(r) nonsymmetric mixtures of four Gaussians: multimodal, transitional and unimodal. See Fig. 5 of [22] for the shapes of the probability density functions.

4. The values of the observed variables x_i were generated according to the LiNGAM (2). Finally, we randomly permuted the order of x_i.

Further we similarly generated 5 datasets based on dense (full) networks, *i.e.*, full DAGs with every pair of variables is connected by a directed edge, under each combination of number of variables p and sample size n. Then we tested DirectLiNGAM and ICA-LiNGAM on the datasets generated by sparse networks or dense (full) networks. For ICA-LiNGAM, the maximum number of iterations was taken as 1000 [6]. The experiments were conducted on a standard PC using Matlab 7.9. Matlab implementations of the two methods are available on the web:
DirectLiNGAM: `http://www.ar.sanken.osaka-u.ac.jp/~inazumi/dlingam.html`
ICA-LiNGAM: `http://www.cs.helsinki.fi/group/neuroinf/lingam/`.

We computed the distance between the true $\mathbf{B}$ and ones estimated by DirectLiNGAM and ICA-LiNGAM using the Frobenius norm defined as

$$\sqrt{\mathrm{trace}\{(\mathbf{B}_{true} - \widehat{\mathbf{B}})^T(\mathbf{B}_{true} - \widehat{\mathbf{B}})\}}. \tag{18}$$

Table 1: Median distances (Frobenius norms) between true **B** and estimated **B** of DirectLiNGAM and ICA-LiNGAM with five replications.

Sparse networks		Sample size		
		500	1000	2000
DirectLiNGAM	dim. = 10	0.48	0.31	0.21
	dim. = 20	1.19	0.70	0.50
	dim. = 50	2.57	1.82	1.40
	dim. = 100	5.75	4.61	2.35
ICA-LiNGAM	dim. = 10	3.01	0.74	0.65
	dim. = 20	9.68	3.00	2.06
	dim. = 50	20.61	20.23	12.91
	dim. = 100	40.77	43.74	36.52
DirectLiNGAM with	dim. = 10	0.48	0.30	0.24
prior knowledge (50%)	dim. = 20	1.00	0.71	0.49
	dim. = 50	2.47	1.75	1.19
	dim. = 100	4.94	3.89	2.27

Dense (full) networks		Sample size		
		500	1000	2000
DirectLiNGAM	dim. = 10	0.45	0.46	0.20
	dim. = 20	1.46	1.53	1.12
	dim. = 50	4.40	4.57	3.86
	dim. = 100	7.38	6.81	6.19
ICA-LiNGAM	dim. = 10	1.71	2.08	0.39
	dim. = 20	6.70	3.38	1.88
	dim. = 50	17.28	16.66	12.05
	dim. = 100	34.95	34.02	32.02
DirectLiNGAM with	dim. = 10	0.45	0.31	0.19
prior knowledge (50%)	dim. = 20	0.84	0.90	0.41
	dim. = 50	2.48	1.86	1.56
	dim. = 100	4.67	3.60	2.61

Tables 1 and 2 show the median distances (Frobenius norms) and median computational times (CPU times) respectively. In Table 1, DirectLiNGAM was better in distances of **B** and gave more accurate estimates of **B** than ICA-LiNGAM for all of the conditions. In Table 2, the computation amount of DirectLiNGAM was rather larger than ICA-LiNGAM when the sample size was increased. A main bottleneck of computation was the kernel-based independence measure. However, its computation amount can be considered to be still tractable. In fact, the actual elapsed times were approximately one-quarter of their CPU times respectively probably because the CPU had four cores. Interestingly, the CPU time of ICA-LiNGAM actually decreased with increased sample size in

Table 2: Median computational times (CPU times) of DirectLiNGAM and ICA-LiNGAM with five replications.

Sparse networks		Sample size		
		500	1000	2000
DirectLiNGAM	dim. = 10	15.16 sec.	37.21 sec.	66.75 sec.
	dim. = 20	1.56 min.	5.75 min.	17.22 min.
	dim. = 50	16.25 min.	1.34 hrs.	2.70 hrs.
	dim. = 100	2.35 hrs.	21.17 hrs.	19.90 hrs.
ICA-LiNGAM	dim. = 10	0.73 sec.	0.41 sec.	0.28 sec.
	dim. = 20	5.40 sec.	2.45 sec.	1.14 sec.
	dim. = 50	14.49 sec.	21.47 sec.	32.03 sec.
	dim. = 100	46.32 sec.	58.02 sec.	1.16 min.
DirectLiNGAM with	dim. = 10	4.13 sec.	17.75 sec.	30.95 sec.
prior knowledge (50%)	dim. = 20	28.02 sec.	1.64 min.	4.98 min.
	dim. = 50	7.62 min.	28.89 min.	1.09 hrs.
	dim. = 100	48.28 min.	1.84 hrs.	7.51 hrs.

Dense (full) networks		Sample size		
		500	1000	2000
DirectLiNGAM	dim. = 10	8.05 sec.	24.52 sec.	49.44 sec.
	dim. = 20	1.00 min.	4.23 min.	6.91 min.
	dim. = 50	16.18 min.	1.12 hrs.	1.92 hrs.
	dim. = 100	2.16 hrs.	8.59 hrs.	17.24 hrs.
ICA-LiNGAM	dim. = 10	0.97 sec.	0.34 sec.	0.27 sec.
	dim. = 20	5.35 sec.	1.25 sec.	4.07 sec.
	dim. = 50	15.58 sec.	21.01 sec.	31.57 sec.
	dim. = 100	47.60 sec.	56.57 sec.	1.36 min.
DirectLiNGAM with	dim. = 10	2.67 sec.	5.66 sec.	12.31 sec.
prior knowledge (50%)	dim. = 20	5.02 sec.	31.70 sec.	38.35 sec.
	dim. = 50	46.74 sec.	2.89 min.	5.00 min.
	dim. = 100	3.19 min.	10.44 min.	19.80 min.

some cases. This is presumably due to better convergence properties.

To visualize the estimation results, Figures 1, 2, 3 and 4 give combined scatterplots of the estimated elements of $\mathbf{B}$ of DirectLiNGAM and ICA-LiNGAM versus the true ones for sparse networks and dense (full) networks respectively. The different plots correspond to different numbers of variables and different sample sizes, where each plot combines the data for different adjacency matrices $\mathbf{B}$ and 18 different distributions of the external influences $p(e_i)$. We can see that DirectLiNGAM worked well and better than ICA-LiNGAM, as evidenced by the grouping of the data points onto the main diagonal.

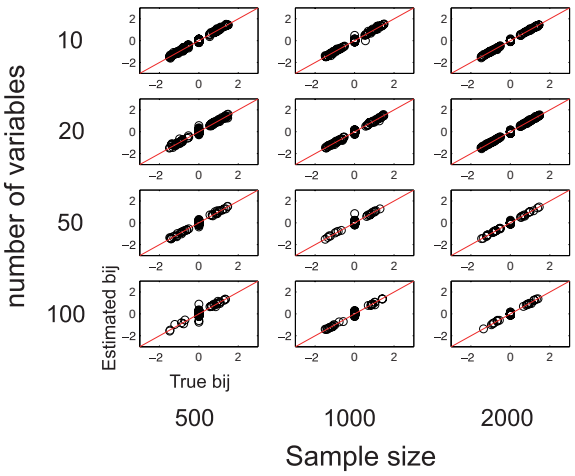

Figure 1: Scatterplots of the estimated b_{ij} by DirectLiNGAM versus the true values for *sparse* networks.

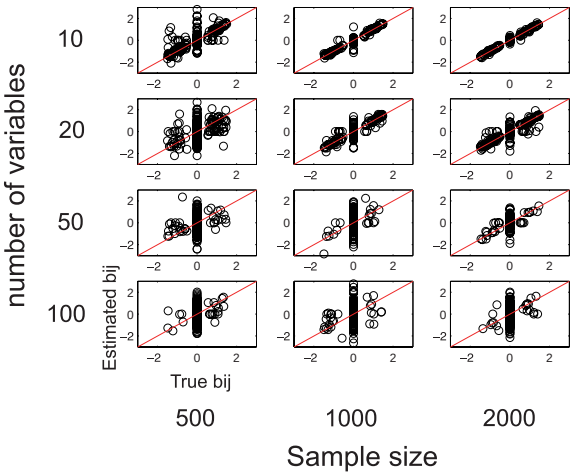

Figure 2: Scatterplots of the estimated b_{ij} by ICA-LiNGAM versus the true values for *sparse* networks.

Finally, we generated datasets in the same manner as above and gave some prior knowledge to DirectLiNGAM by creating prior knowledge matrices $\mathbf{A}^{\mathrm{knw}}$ as follows. We first replaced every non-zero element by unity and every diagonal element by zero in $\mathbf{A}=(\mathbf{I}- B)^{-1}$ and subsequently hid each of the off-diagonal elements, *i.e.*, replaced it by -1, with probability 0.5. The bottoms of Tables 1 and 2 show the median distances and median computational times. It was em-

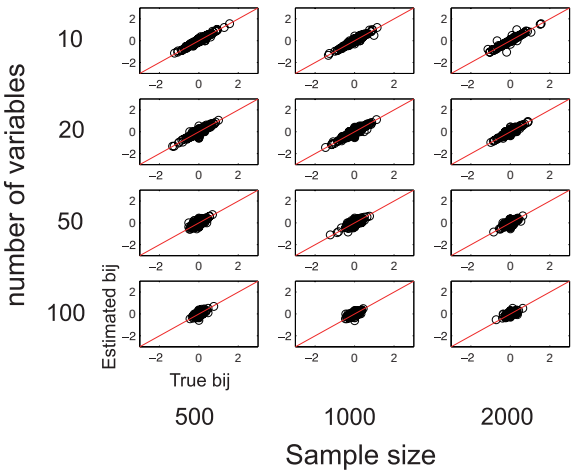

Figure 3: Scatterplots of the estimated b_{ij} by DirectLiNGAM versus the true values for *dense* (full) networks.

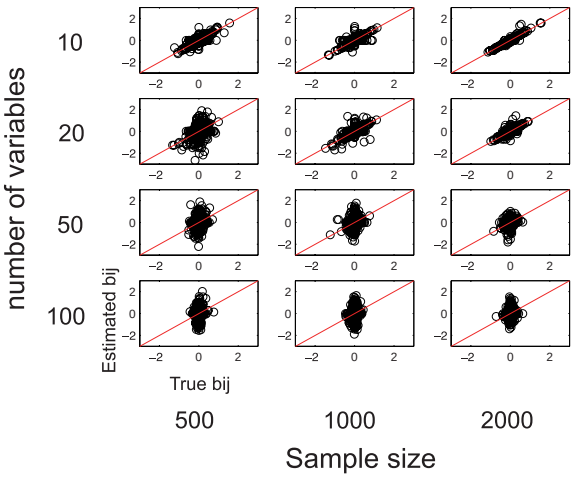

Figure 4: Scatterplots of the estimated b_{ij} by ICA-LiNGAM versus the true values for *dense* (full) networks.

pirically confirmed that use of prior knowledge gave more accurate estimates and less computational times in most cases especially for dense (full) networks. The reason would probably be that for dense (full) networks more prior knowledge about where directed paths exist were likely to be given and it narrowed down the search space more efficiently.

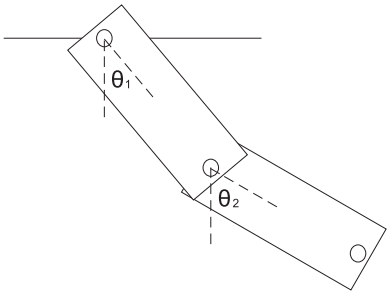

Figure 5: Abstract model of the double-pendulum used in [29].

5 Applications to real-world data

We here apply DirectLiNGAM and ICA-LiNGAM on real-world physics and sociology data. Both DirectLiNGAM and ICA-LiNGAM estimate a causal ordering of variables and provide a full DAG. Then we have two options to do further analysis [9]: i) Find significant directed edges or direct causal effects b_{ij} and significant total causal effects a_{ij} with $\mathbf{A}=(\mathbf{I}-\mathbf{B})^{-1}$; ii) Estimate redundant directed edges to find the underlying DAG. We demonstrate an example of the former in Subsection 5.1 and that of the latter in Subsection 5.2.

5.1 Application to physical data

We applied DirectLiNGAM and ICA-LiNGAM on a dataset created from a physical system called a double-pendulum, a pendulum with another pendulum attached to its end [28] as in Fig. 5. The dataset was first used in [29]. The raw data consisted of four time series provided by Ibaraki University (Japan) filming the pendulum system with a high-speed video camera at every 0.01 second for 20.3 seconds and then reading out the position using an image analysis software. The four variables were θ_1: the angle between the top limb and the vertical, θ_2: the angle between the bottom limb and the vertical, ω_1: the angular speed of θ_1 or $\dot{\theta}_1$ and ω_2: the angular speed of θ_2 or $\dot{\theta}_2$. The number of time points was 2035. The dataset is available on the web:
`http://www.ar.sanken.osaka-u.ac.jp/~inazumi/data/furiko.html`

In [29], some theoretical considerations based on the domain knowledge implied that the angle speeds ω_1 and ω_2 are mainly determined by the angles θ_1 and θ_2 in both cases where the swing of the pendulum is sufficiently small $(\theta_1, \theta_2 \approx 0)$ and where the swing is not very small. Further, in practice, it was reasonable to assume that there were no latent confounders [29].

As a preprocessing, we first removed the time dependency from the raw data using the ARMA (AutoRegressive Moving Average) model with 2 autoregressive terms and 5 moving average terms following [29]. Then we applied DirectLiNGAM and ICA-LiNGAM on the preprocessed data. The estimated

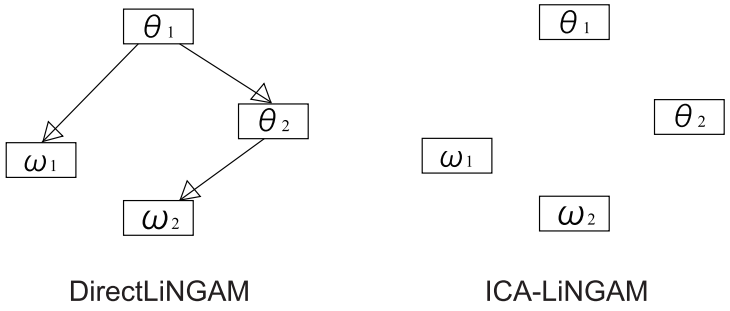

Figure 6: Left: The estimated network by DirectLiNGAM. Only significant directed edges are shown with 5% significance level. Right: The estimated network by ICA-LiNGAM. No significant directed edges were found with 5% significance level.

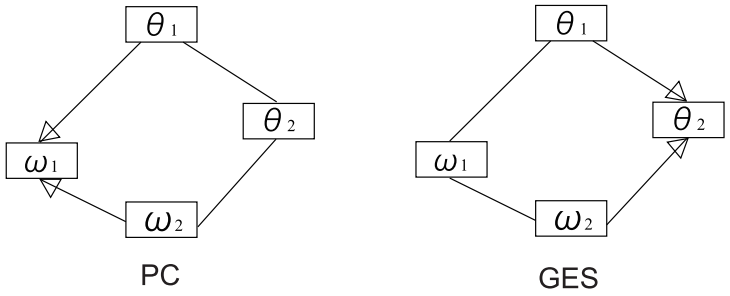

Figure 7: Left: The estimated network by PC algorithm with 5% significance level. Right: The estimated network by GES. An undirected edge between two variables means that there is a directed edge from a variable to the other or the reverse.

adjacency matrices $\mathbf{B}$ of θ_1, θ_2, ω_1 and ω_2 were as follows:

$$\text{DirectLiNGAM} \quad : \quad \begin{array}{c} \\ \theta_1 \\ \theta_2 \\ \omega_1 \\ \omega_2 \end{array} \begin{array}{c} \begin{array}{cccc} \theta_1 & \theta_2 & \omega_1 & \omega_2 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ -0.23 & 0 & 0 & 0 \\ 90.39 & -2.88 & 0 & 0 \\ 5.65 & 94.64 & -0.11 & 0 \end{pmatrix} \end{array}, \tag{19}$$

$$\text{ICA} - \text{LiNGAM} \quad : \quad \begin{array}{c} \\ \theta_1 \\ \theta_2 \\ \omega_1 \\ \omega_2 \end{array} \begin{array}{c} \begin{array}{cccc} \theta_1 & \theta_2 & \omega_1 & \omega_2 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1.45 & 0 & 0 & 0 \\ 108.82 & -52.73 & 0 & 0 \\ 216.26 & 112.50 & -1.89 & 0 \end{pmatrix} \end{array}. \tag{20}$$

The estimated orderings by DirectLiNGAM and ICA-LiNGAM were identical,

but the estimated connection strengths were very different. We further computed their 95% confidence intervals by using bootstrapping [30] with the number of bootstrap replicates 10000. The estimated networks by DirectLiNGAM and ICA-LiNGAM are graphically shown in Fig. 6, where only significant directed edges (direct causal effects) b_{ij} are shown with 5% significance level.[3] DirectLiNGAM found that the angle speeds ω_1 and ω_2 were determined by the angles θ_1 or θ_2, which was consistent with the domain knowledge. Though the directed edge from θ_1 to θ_2 might be a bit difficult to interpret, the effect of θ_1 on θ_2 was estimated to be negligible since the coefficient of determination [1] of θ_2, *i.e.*, $1-\mathrm{var}(\hat{e}_2)/\mathrm{var}(\hat{\theta}_2)$, was very small and was 0.01. (The coefficient of determination of ω_1 and that of ω_2 were 0.46 and 0.49 respectively.) On the other hand, ICA-LiNGAM could not find any significant directed edges since it gave very different estimates for different bootstrap samples.

For further comparison, we also tested two conventional methods [31, 32] based on conditional independences. Fig. 7 shows the estimated networks by PC algorithm [31] with 5% significance level and GES [32] with the Gaussianity assumption. We used the Tetrad IV[4] to run the two methods. PC algorithm found the same directed edge from θ_1 on ω_1 as DirectLiNGAM did but did not found the directed edge from θ_2 on ω_2. GES found the same directed edge from θ_1 on θ_2 as DirectLiNGAM did but did not find that the angle speeds ω_1 and ω_2 were determined by the angles θ_1 or θ_2.

We also computed the 95% confidence intervals of the total causal effects a_{ij} using bootstrap. DirectLiNGAM found significant total causal effects from θ_1 on θ_2, from θ_1 on ω_1, from θ_1 on ω_2, from θ_2 on ω_1, and from θ_2 on ω_2. These significant total effects would also be reasonable based on similar arguments. ICA-LiNGAM only found a significant total causal effect from θ_2 on ω_2.

Overall, although the four variables θ_1, θ_2, ω_1 and ω_2 are likely to be nonlinearly related according to the domain knowledge [28, 29], DirectLiNGAM gave interesting results in this example.

5.2 Application to sociology data

We analyzed a dataset taken from a sociological data repository on the Internet called General Social Survey (`http://www.norc.org/GSS+Website/`). The data consisted of six observed variables, x_1: father's occupation level, x_2: son's income, x_3: father's education, x_4: son's occupation level, x_5: son's education, x_6: number of siblings. The sample selection was conducted based on the following criteria: i) non-farm background (based on two measures of father's occupation); ii) ages 35 to 44; iii) white; iv) male; v) in the labor force at the time of the survey; vi) not missing data for any of the covariates; vii) years 1972-2006. The sample size was 1380. Fig. 8 shows domain knowledge about their causal relations [33]. As shown in the figure, there could be some latent confounders between x_1 and x_3, x_1 and x_6, or x_3 and x_6. An objective of this

[3] The issue of multiple comparisons arises in this context, which we would like to study in future work.

[4] `http://www.phil.cmu.edu/projects/tetrad/`

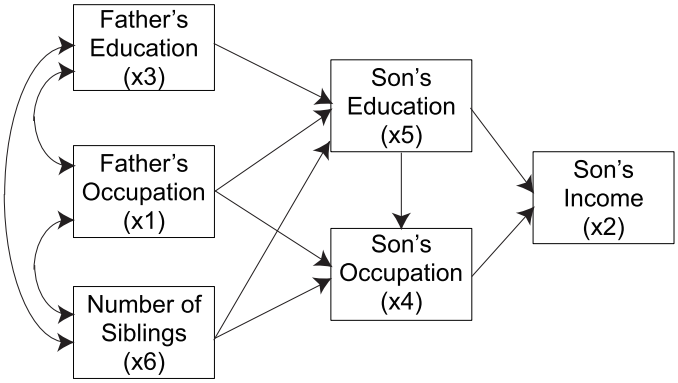

Figure 8: Status attainment model based on domain knowledge [33]. A bi-directed edge between two variables means that the relation is not modeled. For instance, there could be latent confounders between the two, there could be a directed edge between the two, or the two could be independent.

example was to see how our method behaves when such a model assumption of LiNGAM could be violated that there is no latent confounder.

The estimated adjacency matrices $\mathbf{B}$ by DirectLiNGAM and ICA-LiNGAM were as follows:

$$\text{DirectLiNGAM}: \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{pmatrix} 0 & 0 & 3.19 & 0.10 & 0.41 & 0.21 \\ 33.48 & 0 & 452.84 & 422.87 & 1645.45 & 347.96 \\ 0 & 0 & 0 & 0 & 0.55 & -0.18 \\ 0 & 0 & 0.17 & 0 & 4.61 & -0.19 \\ 0 & 0 & 0 & 0 & 0 & -0.12 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}, \quad (21)$$

$$\text{ICA}-\text{LiNGAM}: \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{pmatrix} 0 & 0 & 0.93 & 0 & -0.68 & -0.20 \\ 50.70 & 0 & -31.82 & 200.84 & 65.63 & 336.04 \\ 0 & 0 & 0 & 0 & 0.24 & -0.27 \\ 0.17 & 0 & -0.40 & 0 & -0.14 & -0.14 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.08 & 0 \end{pmatrix} \end{array}. \quad (22)$$

We subsequently pruned redundant directed edges b_{ij} in the full DAGs by repeatedly applying a sparse method called Adaptive Lasso [34] on each variable and its potential parents. See Appendix A for some more details of Adaptive Lasso. We used a matlab implementation in [35] to run the Lasso. Then we

obtained the following pruned adjacency matrices $\mathbf{B}$:

$$\text{DirectLiNGAM} \quad : \quad \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{pmatrix} 0 & 0 & 3.19 & 0 & 0 & 0 \\ 0 & 0 & 0 & 422.87 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.55 & 0 \\ 0 & 0 & 0 & 0 & 4.61 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.12 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}, \tag{23}$$

$$\text{ICA} - \text{LiNGAM} \quad : \quad \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{pmatrix} 0 & 0 & 0.93 & 0 & 0 & 0 \\ 0 & 0 & 0 & 200.84 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.24 & 0 \\ 0 & 0 & 0 & 0 & -0.14 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.08 & 0 \end{pmatrix} \end{array}. \tag{24}$$

The estimated networks by DirectLiNGAM and ICA-LiNGAM are graphically shown in Fig. 9 and Fig. 10 respectively. All the directed edges estimated by DirectLiNGAM were reasonable to the domain knowledge other than the directed edge from x_5: son's education to x_3: father's education. Since the sample size was large and yet the estimated model was not fully correct, the mistake on the directed edge between x_5 and x_3 might imply that some model assumptions might be more or less violated in the data. ICA-LiNGAM gave a similar estimated network but did one more mistake that x_6: number of siblings is determined by x_5: son's education.

Further, Fig. 11 and Fig. 12 show the estimated networks by PC algorithm with 5% significance level and GES with the Gaussianity assumption. Both of the conventional methods did not find the directions of many edges. The two conventional methods found a reasonable direction of the edge between x_1: father's occupation and x_3: father's education, but they gave a wrong direction of the edge between x_1: father's occupation and x_4: son's occupation.

6 Conclusion

We presented a new estimation algorithm for the LiNGAM that has guaranteed convergence to the right solution in a fixed number of steps if the data strictly follows the model and known computational complexity unlike most ICA methods. This is the first algorithm specialized to estimate the LiNGAM. Simulations implied that the new method often provides better statistical performance than a state of the art method based on ICA. In real-world applications to physics and sociology, promising results were obtained. Future works would include i) assessment of practical performance of statistical tests to detect violations of the model assumptions including tests of independence [36]; ii) implementation issues of our algorithm to improve the practical computational efficiency.

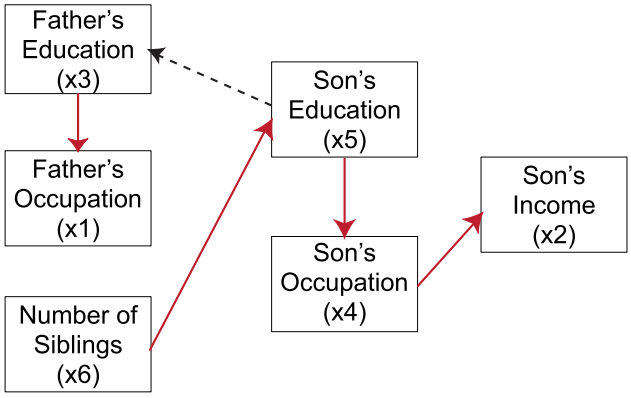

Figure 9: The estimated network by DirectLiNGAM and Adaptive Lasso. A red solid directed edge is reasonable to the domain knowledge.

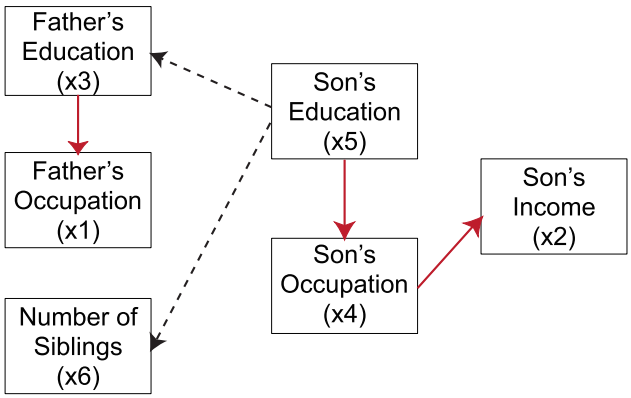

Figure 10: The estimated network by ICA-LiNGAM and Adaptive Lasso. A red solid directed edge is reasonable to the domain knowledge.

Acknowledgements

We are very grateful to Hiroshi Hasegawa (College of Science, Ibaraki University, Japan) for providing the physics data and Satoshi Hara and Ayumu Yamaoka for interesting discussion. This work was partially carried out at Department of Mathematical and Computing Sciences and Department of Computer Science, Tokyo Institute of Technology, Japan. S.S., Y.K. and T.W. were partially supported by MEXT Grant-in-Aid for Young Scientists #21700302, by JSPS Grant-in-Aid for Young Scientists #20800019 and by Grant-in-Aid for Scientific Research (A) #19200013, respectively. S.S. and Y.K. were partially supported by JSPS Global COE program 'Computationism as a Foundation for the Sciences'. A.H. was partially supported by the Academy of Finland Centre of Excellence for Algorithmic Data Analysis.

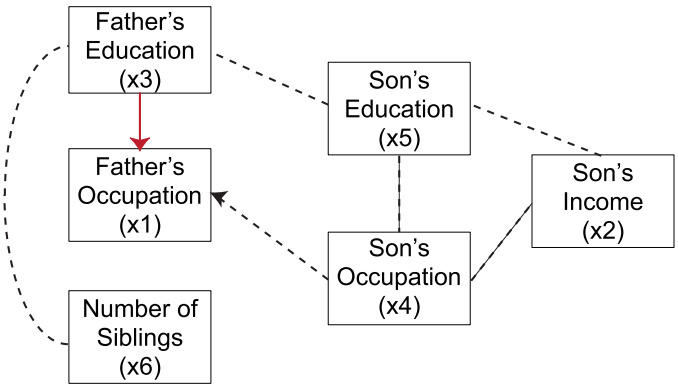

Figure 11: The estimated network by PC algorithm with 5% significance level. An undirected edge between two variables means that there is a directed edge from a variable to the other or the reverse. A red solid directed edge is reasonable to the domain knowledge.

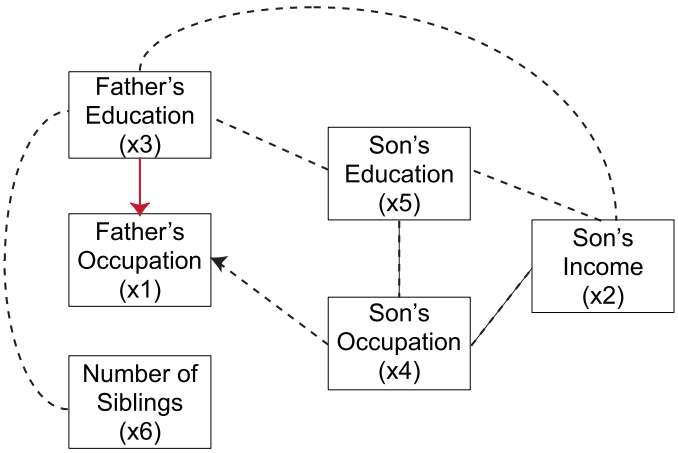

Figure 12: The estimated network by GES. An undirected edge between two variables means that there is a directed edge from a variable to the other or the reverse. A red solid directed edge is reasonable to the domain knowledge.

A Adaptive Lasso

We very briefly review the adaptive Lasso [34], which is a variant of the Lasso [37]. See [34] for more details. The adaptive Lasso is a regularization technique for variable selection and assumes the same data generating process as LiNGAM:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i. \tag{25}$$

A big difference is that the adaptive Lasso assumes that the set of such potential parent variables x_j that $k(j)<k(i)$ is known and LiNGAM estimates the set of such variables. The adaptive Lasso penalizes connection strengths b_{ij} in L_1 penalty by minimizing the objective function defined as:

$$\left\| x_i - \sum_{k(j)<k(i)} b_{ij} x_j \right\|^2 + \lambda \sum_{k(j)<k(i)} \frac{|b_{ij}|}{|\hat{b}_{ij}|^\gamma}, \tag{26}$$

where λ and γ are tuning parameters and $\hat{b}_{ij}$ is a consistent estimate of b_{ij}. In [34], it was suggested to select the tuning parameters by five-fold cross validation and to obtain $\hat{b}_{ij}$ by ordinary least squares regression. The adaptive Lasso has a very attractive property that it asymptotically selects the right set of such variables x_j that b_{ij} is not zero, where $k(j)<k(i)$.

References

[1] Bollen, K.A.: Structural Equations with Latent Variables. John Wiley & Sons (1989)

[2] Rhein, R.O., Strimmer, K.: From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology **1** (2007) 1–37

[3] Londei, A., D'Ausilio, A., Basso, D., Belardinelli, M.O.: A new method for detecting causality in fMRI data of cognitive processing. Cognitive processing **7**(1) (March 2006) 42–52

[4] Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000) (2nd ed. 2009).

[5] Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. Springer Verlag (1993) (2nd ed. MIT Press 2000).

[6] Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research **7** (2006) 2003–2030

[7] Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. Wiley, New York (2001)

[8] Hoyer, P.O., Janzing, D., Mooij, J., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 21. (2009) 689–696

[9] Hyvärinen, A., Zhang, K., Shimizu, S., Hoyer, P.O.: Estimation of a structural vector autoregressive model using non-Gaussianity. Journal of Machine Learning Research **11** (May 2010) 17091731

[10] Lacerda, G., Spirtes, P., Ramsey, J., Hoyer, P.O.: Discovering cyclic causal models by independent components analysis. In: Proc. Uncertainty in Artificial Intelligence (UAI2008). (2008) 366–374

[11] Amari, S.: Natural gradient learning works efficiently in learning. Neural Computation **10** (1998) 251–276

[12] Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. on Neural Networks **10** (1999) 626–634

[13] Himberg, J., Hyvärinen, A., Esposito, F.: Validating the independent components of neuroimaging time-series via clustering and visualization. NeuroImage **22** (2004) 1214–1222

[14] Shimizu, S., Hyvärinen, A., Kawahara, Y., Washio, T.: A direct method for estimating a causal ordering in a linear non-gaussian acyclic model. In: Proc. the 25th Conf. on Uncertainty in Artificial Intelligence (UAI2009), Montreal, Canada, AUAI Press (2009) 506–513

[15] Inazumi, T., Shimizu, S., Washio, T.: Use of prior knowledge in a non-Gaussian method for learning linear structural equation models. In: Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA2010). (2010) 221–228

[16] Sogawa, Y., Shimizu, S., Kawahara, Y., Washio, T.: An experimental comparison of linear non-gaussian causal discovery methods and their variants. In: Proc. International Joint Conference on Neural Networks (IJCNN2010). (2010) 768–775

[17] Hoyer, P.O., Shimizu, S., Kerminen, A., Palviainen, M.: Estimation of causal effects using linear non-gaussian causal models with hidden variables. International Journal of Approximate Reasoning **49**(2) (2008) 362–378

[18] Comon, P.: Independent component analysis, a new concept? Signal Processing **36** (1994) 62–83

[19] Eriksson, J., Koivunen, V.: Identifiability, separability, and uniqueness of linear ICA models. IEEE Signal Processing Letters **11** (2004) 601–604

[20] Darmois, G.: Analyse générale des liaisons stochastiques. Rev. Inst. Internat. Stat. **21** (1953) 2–8

[21] Skitovitch, W.P.: On a property of the normal distribution. Doklady Akademii Nauk SSSR **89** (1953) 217–219

[22] Bach, F.R., Jordan, M.I.: Kernel independent component analysis. Journal of Machine Learning Research **3** (2002) 1–48

[23] Gretton, A., Bousquet, O., Smola, A.J., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: Algorithmic Learning Theory: 16th Int. Conf. (ALT2005). (2005) 63–77

[24] Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical Review E **69**(6) (2004) 066138

[25] Hyvärinen, A.: New approximations of differential entropy for independent component analysis and projection pursuit. In: Advances in Neural Information Processing Systems (NIPS1997). Volume 10. (1998) 273–279

[26] Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research **8** (2007) 613–636

[27] Silva, R., Scheines, R., Glymour, C., Spirtes, P.: Learning the structure of linear latent variable models. Journal of Machine Learning Research **7** (Feb 2006) 191–246

[28] Meirovitch, L.: Elements of Vibration Analysis (2nd ed.). McGraw-Hill (1986)

[29] Kawahara, Y., Shimizu, S., Washio, T.: Analyzing relationships among ARMA processes based on non-Gaussianity of external influences. Neurocomputing (2011) Conditionally accepted.

[30] Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)

[31] Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review **9** (1991) 67–72

[32] Chickering, D.: Optimal structure identification with greedy search. Journal of Machine Learning Research **3** (2002) 507–554

[33] Duncan, O.D., Featherman, D.L., Duncan, B.: Socioeconomic Background and Achievement. Seminar Press, New York (1972)

[34] Zou, H.: The adaptive Lasso and its oracle properties. Journal of the American Statistical Association **101** (2006) 1418–1429

[35] Sjöstrand, K.: Matlab implementation of LASSO, LARS, the elastic net and SPCA (June 2005) Version 2.0.

[36] Gretton, A., Györfi, L.: Consistent nonparametric tests of independence. Journal of Machine Learning Research **11** (2010) 1391–1423

[37] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society: Series B **58**(1) (1996) 267–288